LABORatorio R. Revelli
Centre for Employment Studies

Jesus vs. Hillel.
Moral and Social Norms in Heterogeneous Populations

Matteo Richiardi
LABORatorio Riccardo Revelli Centre for Employment Studies

Version: 02/11/2005

Collegio "Carlo Alberto" via Real Collegio, 30 - 10024 Moncalieri (TO)
Tel. +39 011.640.26.59/26.60 - Fax +39.011.647.96.43 – www.labor-torino.it - labor@labor-torino.it
LABOR is an independent research centre within Coripe Piemonte

Abstract

This paper presents an idealized model of social interaction, where preferences are private information and individuals cannot condition their behavior on the identity of whom they are interacting with. An optimal decentralized benchmark rule is identified, where each individual imposes some restriction on what people interacting with him cannot do. Social norms arise in the model as a consequence of reciprocating behavior. I show that social norms can always be reversed, as long as there remains a minimal level of diversity in individual choices. Social norms turn out to be less efficient than democracy as a way to obtain homogeneity in individual behavior. However, both mechanisms are welfare-reducing w.r.t. the decentralized benchmark. Moreover, imposing a single behavior by democratic decision is more welfare-reducing the more fragmented society is (thus the larger the "threat" of invasion from a population with adverse preferences). Unfortunately, this is exactly the case when the law has a higher probability of being implemented. Finally, the democratic decision of banning a specific action is analyzed. I found that bans can be both welfare-reducing or welfare-enhancing w.r.t. the decentralized benchmark. However, they are more likely to be welfare-reducing when they hurt more people (for instance the larger the "threat" of invasion from a population with adverse preferences), but this is when they are more likely to be implemented.

Keywords: Norms, Reciprocity, Liberalism, Socialism, Democracy, Immigration
JEL Classification: D7, D63, D64, P50

> The Talmud tells that a gentile came to Hillel saying that he would convert to Judaism if Hillel could teach him the whole Torah in the time that he could stand on one foot. Hillel converted the gentile by telling him, "That which is hateful to you, do not do to your neighbor. That is the whole Torah; the rest is commentary. Go and study it."
>
> "And seeing the multitudes, Christ went up into a mountain. And when he was set, his disciples came unto him. And he opened his mouth, and taught them, saying – Do unto others as you would have them do unto you." (Matthew 7:12)

1. Introduction

Aristotle wrote that humans are social animals, and indeed, people generally live in societies. People live together because they can benefit from mutual interaction. Of course, they can also suffer from these interactions. However, what is important for the analysis conducted below is that, living in societies, they can hardly escape to interact with each other. No matter how we behave, we affect other people's well being.

The emergence of pro-social behavior in human societies has been the matter of thorough investigations. Two kinds of explanations have been advanced. One builds upon the hypothesis of rational behavior of self-interested individuals, and stresses the importance of reciprocal altruism (Triver, 1971; Axelrod and Hamilton, 1981): individuals cooperate in exchange of other people's cooperation. The other stresses the importance of cultural (Cavalli-Sforza et al., 1981; Boyd and Richerson, 1985) and genetic (Lumsden and Wilson, 1981; Simon, 1983; Wilson and Dugatkin, 1997; Sober&Wilson, 1998) evolution.

In particular, Fehr and Fischbacher (2004) review evidence that human behavior is often based on *conditional* cooperation, i.e. cooperate if other group members cooperate, and defect if other group members defect. They stress the importance of mechanisms such as expectations, reputation and punishment in order to explain the emergence of reciprocal altruism. However, as Gintis (2000) argues, precisely when a group is threatened and is thus most in need of pro-social behavior the probability of future interactions goes down, together with the incentives for reciprocal altruism.

It is no surprise then that many studies have shown[1] that people are not only motivated by economic self-interest but also by norms of fairness and reciprocity, that in turn could be explained in terms of evolutionary selection, as sketched above. Religion is one of the mechanisms for strengthening these social norms.

However, although in many cases it is straightforward to identify what is a pro-social behavior, in general individual preferences are private information. Thus, if player A wants to act in an altruistic way towards player B, player A has to guess which action

[1] see the references in the review paper by Fehr and Fischbacher cited above. In particular Camerer (2003) suggests that retaliation seems often to reflect purely emotional responses to deviations, since it cannot serve any strategic purposes (at least intentionally).

will please the most player B. This point has largely been neglected by the scientific literature, which assumes that the pro-social behavior is always clearly identified. The focus there is on individual costs (and social benefits) of pro-social behavior, and on the incentives to free-ride.

A different approach may complement the analysis. Suppose you do not know what your neighbour likes (you may even not know the identity of the people you are interacting with). Which is the socially optimal rule of behavior? Which is the rule that will prevail under a democratic voting process? Which is the rule that will emerge if everybody will retaliate, i.e. behave to others as others have acted upon them? How do these different rules of behavior compare, for different distributions of preferences in the society? What will happen if a population with different preferences (say, immigrants) joins in? These are the questions that this paper addresses.

Note that this perspective is the same adopted in the religious literature, which generally makes the assumption that, not knowing what your neighbour likes, you should act as if your neighbour were not too different from yourself. This gave rise to a number of "golden rules", of which two prototypes are the Christian and the Jewish golden rule quoted above. The rule stated by Jesus in his Mountain speech (hereafter, J-rule) prescribes to do what you think is good; the rule stated by Hillel (hereafter, H-rule) prescribes not to do what you think is bad. In the history of philosophy there are many antecedents to both rules. On Jesus side we have the Greek philosophers Sextus, Aristotle, Aristippus and Isocrates, while on Hillel side we have Pittacus and Thales and the Chinese philosopher Confucius.

Of course, one could question the assumption that individual preferences are private information, arguing that in general people can and do communicate. However, the fact itself that many philosophers implicitly made the same assumption points to the relevance of social situations in which individuals cannot or do not want to communicate, or do not want to please others irrespective of their own beliefs about what is and what is not right / good. This last possibility deserves some more comments. It is interesting that a moral norm may value the general principle of pleasing / not harming others, given that other people's preferences may be very different from those endorsed by the same moral norm, while stressing the superiority of those endorsed preferences. The moral norm would then point to a sort of trade-off between consenting to do what other people like[2], and "showing" them what they *should* like. Of the two golden rules stated above, the H-rule pushes the balance more in the direction of pleasing others, while the J-rule offers a corner solution to the above trade-off. [3] It is

[2] the most pro-social rule in a world with perfect knowledge would simply be "do others what they like", or "let them choose what to do, without restrictions"

[3] Note however that even the J-rule is stated in terms of fairness and reciprocity (it should otherwise read something like "do others what you think is right").

therefore easy to find a flavour of socialism in the J-rule, while the H-rule looks definitely more liberal. [4]

The purpose of the first part of the paper is to investigate their implications for aggregate welfare in the simplest possible model. One version of the H-rule is then used as a benchmark in the remaining part of the paper, when retaliation, voting and immigration are considered. The model is described in section 2. Section 3 presents the J-rule and one version of the H-rule. Section 4 presents another version of the H-rule that turns out to be the optimal decentralized interaction rule, given the assumptions that interaction must entail some restrictions on individual behaviour, and that these restrictions cannot be tailored on the identity of the interacting partner. This rule is then used as a benchmark in the remaining part of the paper. Section 5 investigates what happens when individuals depart from the moral norm and play a retaliation (tit-for-tat) strategy ("what has been done unto you, do it to others"), which as we have seen is after all a very common behavior. [5] Conformity in behaviour is obtained within groups, while heterogeneity can still emerge between groups, two characteristics that remind of the establishment of social norms. Section 6 considers the incentives and implications to impose a single behavior, the one preferred by the relative majority, by law (hence the name *Democratic Jesus*, or DJ-rule), while section 7 considers the incentives and implications to forbid a particular behavior, the one hasted by the relative majority, by law (hence the name *Democratic Hillel*, or DH-rule). These extensions are discussed with explicit reference to the possibility of a dynamic change in the composition of the population, and thus in the distribution of preferences, for instance due to immigration. Section 8 summarizes and concludes.

2. The model

There are N individuals, who can be in 3 different states (call them *Left*, *Center* and *Right*), which they can choose. Individuals have preferences over their states: they love one state, they are neutral with respect to another state and they hate the remaining state. This identifies only 6 possible combinations. Denote with $p_1 \ldots p_6$ the fractions of the population characterized by each combination of preferences, as in table 1. That is, drawing randomly one individual, she will be of type i with probability p_i.

[4] The reader should not consider the results of this paper as a judgement over different religious prescriptions. The two behavioral rules considered are named after Jesus and Hillel for ease of identification, but many other references, aside all the philosophers cited above, could be found.

[5] This rule has also noble origins, reminding the "eye for eye, tooth for tooth" prescription of the Bible (Exodus 21:18-19, 22-25, and Leviticus 24:17-21) and an almost identical prescription to be found in the Hammurabi code (8^{th} century B.C.). However, as it will be clear later, the "tit-for-tat" rule used in the paper doesn't allow to address the reaction specifically to the offender, thus the label BT, for "blind tit-for-tat", that will be used.

Type	Loved state	Hated state	Share
1	Left	Center	p_1
2	Left	Right	p_2
3	Center	Left	p_3
4	Center	Right	p_4
5	Right	Left	p_5
6	Right	Center	p_6

Table 1: Distribution of preferences in the population

Left alone, anyone would choose her preferred state. However, individuals have to interact. Interaction always involves one *active* and one *passive* player[6]. When two persons meet, the active player may impose some restrictions on the passive player's choice. In such a simple model, restrictions may be of two forms only: *weak* restrictions, where only one choice is forbidden, and *strong* restrictions, where two out of three choices are forbidden (this is equivalent to imposing the remaining choice). After each interaction, the passive player gets a payoff of +1 if she is in her loved state, a payoff of 0 if she is in her neutral state, and a payoff of –1 if she is in her hated state. The active player does not get any feedback[7]. Aggregate welfare is defined only in terms of the mean π of the payoffs. [8]

Note that there is no *strategic* interaction in the model: the passive player's payoff depends on the active player's choice, but the active player's choice does not depend on the passive player's action in any way. This implies that game-theoretic solution concepts like Nash equilibrium become useless.

Note also that individuals would be better-off if they lived alone. I do not model here the benefits from living within a society: simply suppose that they are big enough to prevent people from running away. In such a simplified setting, it is easy to see that weak restrictions should be preferred to strong restrictions. I first focus on strong restrictions, and allow weak restrictions only from section 4 onwards.

3. Strong restrictions

3.1 The J-rule

Transposing Jesus' golden rule in the model is quite straightforward: the active player always imposes to the passive player his preferred action. Suppose two individuals, A and B, meet. Player A is the active one. He hates *Left* and loves *Right* (he is thus neutral

[6] Agents can play both roles interchangeably.

[7] We can alternatively suppose that he receives a positive payoff deriving from acting accordingly to his moral norm, or that he receives a positive (negative) payoff in case the opponent is in some particular state. This will not change qualitatively any of the results. See below for further discussion.

[8] In a companion paper, I define aggregate welfare both in terms of the average payoff and in terms of the payoff variance (as a measure if inequality), and compare in more detail the J-rule with the first operalization of the H-rule proposed here.

with respect to *Center*). Player B is the passive one. She loves *Left* and hates *Right* (she is neutral with respect to *Center*, like player A). Suppose A follows the J-rule. He will play *Right*, setting B's state to *Right*. B will then have a payoff of –1. A will get no payoff.[9]

It is straightforward to see that when all individuals share the same preferences (*polarization*) the J-rule works wonders, and the expected payoff from any interaction is 1. In the other extreme case, when preferences are equally distributed in the population (*fragmentation*) and $p_1 = p_2 = \ldots = p_6 = {}^1/_6$, it is again straightforward to see that the expected payoff is 0. However, it turns out to be possible to have even negative expected payoffs, for particular distributions of preferences.

To derive the expected payoffs from one random interaction, and thus the average payoff, consider an active player of type 1 (he loves L*eft* and hates *Center*), who meets in turn all other (passive) individuals, including himself. If he follows the J-rule, he will play *Left*, causing a payoff of +1 in $(p_1+p_2)N$ individuals, and a payoff of –1 in $(p_3+p_5)N$ individuals. Note that there are $(p_1+p_2)N$ individuals like him in the population.
Suppose now that everybody meets everybody else both as active and as passive player. The average payoff is then

$$\pi_J = (p_1+p_2)(p_1+p_2-p_3-p_5)+(p_3+p_4)(-p_1+p_3+p_4-p_6)+(p_5+p_6)(-p_2-p_4+p_5+p_6) \qquad (1)$$

By numerically evaluating this function for all possible combinations of preferences obtained by changing any p_i at a time with a step equal to .05 we obtain the distribution shown in Figure 1.

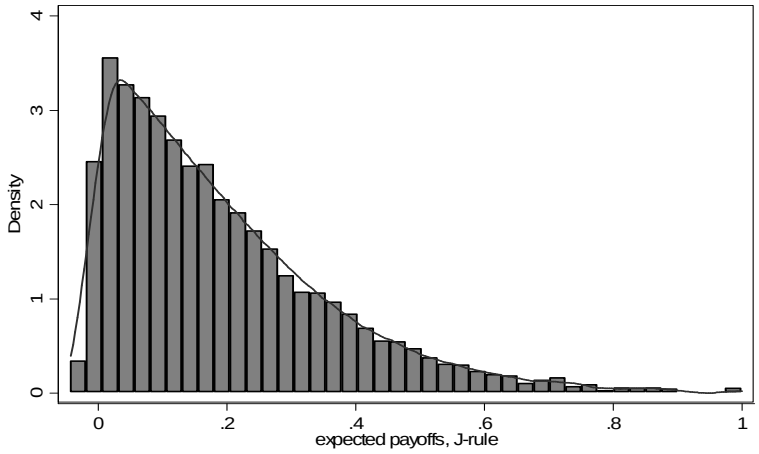

Figure 1: Expected payoff of the J-rule, entire spectrum of preferences

[9] Alternatively, suppose the active player has a preference not only on his state, but also on other people's state (although weaker). He will thus get a payoff of $\alpha < 1$ after imposing his preferred choice at each interaction.

3.2 The H-rule

The operationalization of the H-rule turns out to be more arbitrary. Sticking to the idea of modelling strong restrictions, we consider that under the H-rule, being the action corresponding to the hated state banned, the active player randomises between the actions corresponding to his loved and neutral state. In the example above, A would then randomise between *Right* and *Center*, causing a payoff for B either equal to -1 or equal to 0. Note that in the case of extreme polarization ($p_i = 1$; $p_{-i} = 0$), the expected payoff from each random interaction is 0.5, while in the case of extreme fragmentation ($p_i = p_j \forall i, j$) the expected payoff is equal to 0, and is thus equivalent to the expected payoff from the J-rule.

With an argument similar to the one proposed for the J-rule it is possible to obtain the expression for the expected payoff:

$$\pi_H = \frac{1}{2}\begin{pmatrix}(p_1+p_6)(p_1+p_6-p_3-p_4)+\\(p_2+p_4)(p_2+p_4-p_5-p_6)+\\(p_3+p_5)(p_3+p_5-p_1-p_2)\end{pmatrix} \qquad (2)$$

In Richiardi (2005) the two expressions for the J-rule and the H-rule are compared, and the following conclusions reached:

- which rule gives a higher expected payoff depends on the distribution of preferences in the population;
- when the preferences are highly fragmented, the two rules give roughly the same expected payoff;
- as the preferences become more polarized, the fraction of combinations favourable to the J-rule increase, and reaches 100% when the preferences are totally polarized (note that there are 6 ways of obtaining totally polarized preferences).

So, the J-rule turns out to stochastically dominate the H-rule, for an unknown distribution of preferences.

4. Weak restrictions (H-rule revisited)

As already noted however, there are other ways to interpret the Hillel rule, in the simple setting of this model. A possibility (that turns out to be very close in letter and spirit to the original formulation by Hillel) is to let player B choose whatever action she prefers, as long as it does not correspond to the hated state of player A. The active player simply sets a ban on what he doesn't like. [10]

It is trivial to show that in this case the expected payoff is always greater than zero and is equal to:

[10] This version was suggested by Sorin Solomon.

$$\begin{aligned}\pi_H = \quad & (p_1 + p_6)(p_1 + p_2 + p_5 + p_6) + \\ & (p_2 + p_4)(p_1 + p_2 + p_3 + p_4) + \\ & (p_3 + p_5)(p_3 + p_4 + p_5 + p_6)\end{aligned} \tag{3}$$

It is also straightforward to show that this rule is always superior to the J-rule. Simply consider that for any matching between an active and a passive player, the choice set for the passive player is expanded with the H-rule. She will always be able to choose the active player's preferred choice. However, she will also be allowed to choose the active player's neutral choice. Her payoffs cannot therefore be lower. Given that the unrestricted choice (equivalent to non-interaction) for the passive player is not possible, there is no rule based only on *individual knowledge* (own preferences) better than the H-rule: ban the choice corresponding to what you don't like. I will therefore consider it as a benchmark, to be compared with rules based also on *aggregate knowledge* (for instance, on how many people like or dislike a particular choice).

5. From moral to social norms: the BT-rule

Suppose now that all individuals act according to the following strategy: "if nobody acted to you, play according to the H-rule; otherwise do what your last opponent did to you". This rule is reminiscent of the "tit-for-tat" strategy, with the only difference that the reciprocal behavior cannot be targeted to specific individuals. Thus, retaliation is directed towards society in general. For this reason it will be labelled *Blind tit-for-tat* (BT).

It is easy to see that this strategy always leads to the selection of a single action. [11]
Note that this process of *path dependency* closely resembles the creation of a social norm, which prescribes to play one single action, irrespective of individual preferences. [12] Should we have two distinct populations with the same distribution of preferences, it may happen to observe the selection of a different action within each population, as the social norm of that community. In fact, it is well known that the existence of social norms creates conformity within groups and heterogeneity across groups (Gintis, 2003).

An example of the dynamics that lead to the selection of a social norm is shown in Figure 2. It is interesting to note that the convergence to a given action is not exponential in the share of the population that has already adopted it, as one may have expected. In the example below, for instance, *Left* was played by more than 80%, although it was eventually disregarded.

[11] In Richiardi (2005) I show that by allowing few individuals (as little as 1% of the population) to play according to strong restrictions rules (the J-rule and the first version of the H-rule), the outcomes are almost undistinguishable from those obtained when everybody plays according to these rules. This result doesn't hold for weak restrictions.

[12] Some may argue that the establishment of a social norm involves some modification of individual preferences. Casual observation that individuals often complain about what they regard as an implicit imposition by "the Society" seems to contradict this, showing that preferences might not be completely shaped by social norms, after all. Note however that preferences remain unobservable.

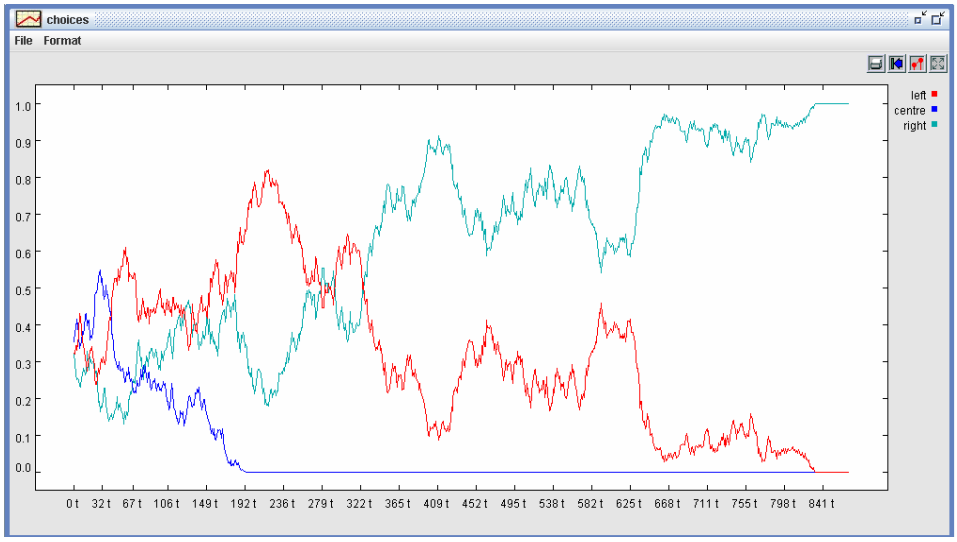

Figure 2: Selection of a social norm (simulation result)

To explain this, suppose that there are only two remaining actions, say α and β, one played by a fraction α of the population, and the other played by the remaining fraction $\beta = (1 - \alpha)$. For α to increase after the next interaction, it must happen that a follower of α is chosen as the active player, and a follower of β is chosen as the passive player (*vice versa* for β to increase):

$$\begin{aligned} prob(\alpha \uparrow) &= \alpha(1-\alpha) \\ prob(\alpha \downarrow) &= (1-\alpha)\alpha \\ prob(\alpha =) &= \alpha^2 + (1-\alpha)^2 \end{aligned} \qquad (4)$$

So, irrespective of how many individuals are playing α, the probability that this share goes up is always equal to the probability that it goes down: the law of motion of α is a random walk. [13]

A similar analysis can be applied from the onset: which action will eventually be selected depends on the initial distribution of preferences in the population and on the (random) order of interactions, and the selection probabilities depend linearly on p_i.

6. Moral norms, social norms and the law: the DJ-rule.

Suppose now that individuals are asked to vote for prescribing one particular action by law. I will label this case *Democratic Jesus* (DJ): "do others what is prescribed by the law". Under the majority rule, the action that will become law is the one preferred by

[13] Note that one single follower of Jesus would be enough to convert a whole population of "blind Hammurabis". However, two followers of the J-rule with different preferences over their most preferred action would lead to never-ending oscillating outcomes.

the relative majority. Suppose for instance, without loss of generality, that $p_1+p_2 > p_3+p_4$ and $p_1+p_2 > p_5+p_6$: *Left* becomes law. The average payoff from each interaction is thus:

$$\pi_L = p_1 + p_2 - p_3 - p_5 \qquad (5)$$

Note that the DJ-rule is always at least as good as any social norm: by construction, the action that is established under the DJ-rule is the one preferred by the relative majority, while under the BT-rule any action can be selected[14].

Note also that Democratic Jesus is always worse that (Anarchic) Hillel. The difference between the expected payoffs under the two rules is:

$$\pi_L - \pi_H = -(p_2 + p_4)(p_3 + p_4) - (p_1 + p_6)(p_5 + p_6) - 2(p_3 + p_5)(1 - p_1 - p_2) \qquad (6)$$

with only negative terms (subscript *L* stands for the case when *Left* is imposed by law).

Moreover, this difference becomes increasingly negative as p_3, p_4, p_5 and p_6 increase, that is the share of those who do not love the imposed choice (*Left*) increases.

Of course, the real winners when *Left* is imposed by law are those who love *Left*. They would get a payoff equal to 1 for each interaction under the law, while they would get an expected payoff equal to $p_1+ p_2+ p_4+ p_6 = 1- p_3 - p_5$ under the H-rule. Thus, their incentive to ask for a vote is

$$\Delta_L = p_3 + p_5 \qquad (7)$$

and is increasing in the fraction of the population who dislikes *Left*.

Now, suppose the distribution of preferences in the society changes, for instance due immigration, in a way that is adverse to the relative majority. In the example above, suppose a sub-population of people who dislike *Left* joins in. The more the "threat of invasion" by this adverse group, the higher is the incentive for the incumbent relative majority to pass a law that prescribes *Left* for everybody, and the more welfare-reducing this turns out to be. [15]

This is shown in Figure 3, where the gains for the relative majority are compared to those for the whole society as the fraction of people who dislike *Left* increases. The figure shows simulation results when p_4 and p_6 are set to zero, and all other shares are changed by steps equal to .05.

[14] although the action preferred by the relative majority is more likely to emerge as the social norm.

[15] Actually, if the proposal to vote for a single action to be prescribed by the law is also agreed upon democratically (*i.e.* by voting), as long as the relative majority is not an absolute majority the incentives for the other groups would be high enough to prevent such a welfare reducing decision.

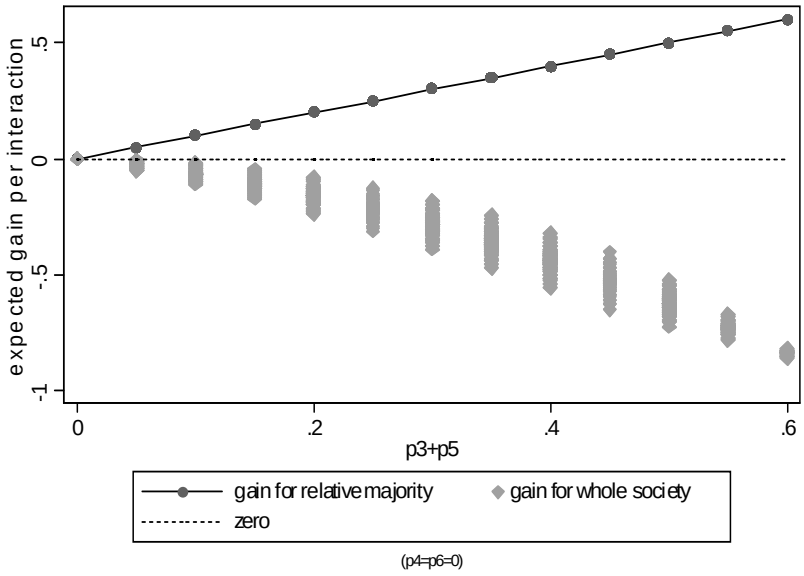

Figure 3: Private and social gains from imposing a single action by law

7. Moral norms, social norms and the law: the DH-rule.

Rather than agreeing on imposing a single action by law, individuals may agree (or democratically forced to agree) on banning one particular action. I will label this case *Democratic Hillel* (DH): "don't others what is forbidden by the law". Under the majority rule, the action that will be banned is the one disliked by the relative majority.

Suppose for instance, without loss of generality, that $p_2+p_4 > p_1+p_6$ and $p_2+p_4 > p_3+p_5$: *Right* becomes outlaw. The average payoff from each interaction is now:

$$\pi_{\bar{R}} = p_1 + p_2 + p_3 + p_4 \tag{8}$$

(where subscript $\bar{R}$ stands for the case when *Right* is banned by law).

The difference between the expected payoffs under the two rules is:

$$\begin{aligned} \pi_{\bar{R}} - \pi_H = \quad & (1-p_2-p_4)(p_1+p_2+p_3+p_4) \\ & -(p_1+p_6)(p_1+p_2+p_5+p_6)-(p_3+p_5)(p_3+p_4+p_5+p_6) \end{aligned} \tag{9}$$

The comparison between $\pi_{\bar{R}}$ and π_H, that is between Democratic Hillel and (Anarchic) Hillel, is thus not straightforward. In facts, the DH-rule can lead both to higher and to lower average payoffs that the H-rule. By numerically evaluating this function for all possible combinations of preferences (obeying to the inequalities above) obtained by changing any p_i at a time with a step equal to .05 we obtain the distribution shown in Figure 4.

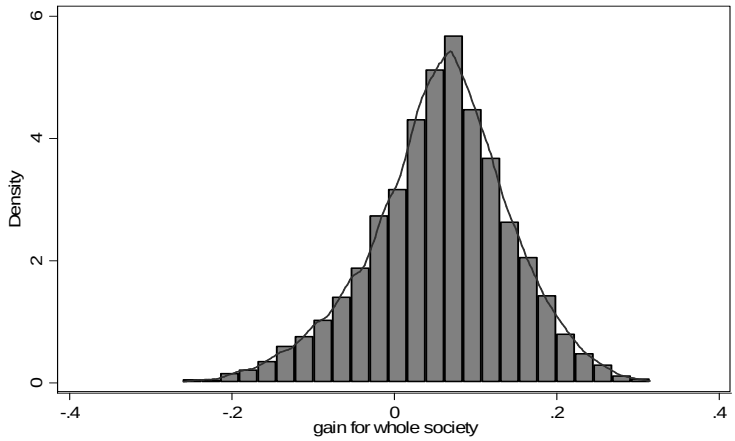

Figure 4: Expected gains from moving from the H-rule to the DH-rule, entire spectrum of preferences

What is important, however, is that the decentralized rule improves over the centralized one as p_5 and p_6 increase[16], that is as the share of those who like the banned action gets bigger.

Again, let's look at the private incentives to establish the ban. Under the DH-rule, whenever *Right* is declared outlaw right-phobics get a payoff of 1 for each interaction. The same right-phobics would get a payoff equal to $p_2(1-p_3-p_5)+p_4(1-p_1-p_6)$ under the H-rule.

Thus, their incentive to ask for a vote is

$$\Delta_{\bar{R}} = 1-p_2(1-p_3-p_5)-p_4(1-p_1-p_6) \tag{10}$$

and is decreasing in p_2 and p_4, *i.e.* in those who dislike *Right*, and increasing in all other types. This result is coherent with what intuition would suggest: if everybody dislikes *Right* there is no need to impose a ban on *Right*. To benefit the most the promoters, the ban must hit a greater number of persons, as in the case when immigrants who love *Right* join in. Unfortunately, this is exactly the case when it hurts the most society as a whole.

This is shown in Figure 5, where the gains for the relative majority are compared to those for the whole society as the fraction of people who dislike Right increases. The figure shows simulation results when p_4 and p_6 are set to zero, and all other shares are changed by steps equal to .05.

[16] The only partial derivatives of eq. 8 with constant sign are those with respect to p_5 and p_6, and they are negative.

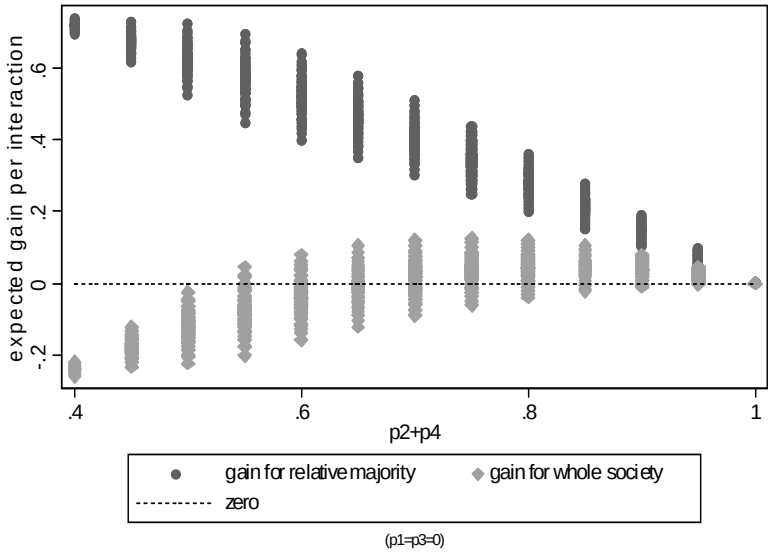

Figure 5: Private and social gains from banning one action by law

8. Summary and conclusions

This paper presents an idealized model of social interaction, where preferences are private information and individuals cannot condition their behavior on the identity of whom they are interacting with. I focus on the costs of social interaction, and abstract altogether from the benefits (which however are assumed to be great enough to prevent people from running away from one another). Given the assumption that interaction must restrict choice (and is therefore welfare reducing), I identify an optimal decentralized rule where each individual imposes some restriction on what people interacting with him cannot do. This is reminiscent of the statement by the Jewish philosopher Hillel (I century B.C.), which was supposed to summarize in only one golden rule the moral message of Hebraism: "don't do others what you don't like them to do to you". This decentralized rule, based only on what individuals know about themselves, is then used as a benchmark in the rest of the paper.

I then turn to the establishment of social norms, that arise as a consequence of reciprocating behavior (where *actions*, rather than *intentions*, are reciprocated). I show that social norms can always be reversed, as long as there remains a minimal level of diversity in individual choices. I then compare social norms with rules based on aggregate knowledge, as discovered by the democratic voting process.

Social norms turn out to be less efficient than voting for the establishment of a single behavior by law, but even the latter is welfare-reducing, w.r.t. the decentralized benchmark of the Hillel rule. Moreover, imposing a single behavior by democratic decision is more welfare-reducing the more fragmented society is (thus the larger the "threat" of invasion from a population with adverse preferences). Unfortunately, this is exactly the case when the law has a higher probability of being implemented.

Finally, the democratic decision of banning a specific action is analyzed. I found that bans can be both welfare-reducing or welfare-enhancing w.r.t. the decentralized benchmark. However, they are more likely to be welfare-reducing when they hurt more people (for instance the larger the "threat" of invasion from a population with adverse preferences), that is when they are also more likely to be implemented.

Acknowledgements: A Lagrange fellowship by ISI Foundation is gratefully acknowledged. This work owns a lot to suggestions by Sorin Solomon, to whom I am greatly indebted. I also wish to thank the audience at ESA 2005 European Regional Meeting in Alessandria, Italy, WEHIA 2005 conference in Essex, Great Britain and Wild@Ace 2004 conference in Torino, Italy for providing feedbacks as this paper developed. In particular, I wish to thank Riccardo Boero, whose critical comments as a discussant of the paper helped in convincing me that I was right. Jennifer Chubinski provided an English language check. All remaining errors and omissions remain naturally my responsibility alone.

References

Axelrod R., Hamilton W.D. (1981), "The evolution of cooperation", *Science*, 211, 1390–1396

Boyd R., Richerson, P.J. (1985), *Culture and the Evolutionary Process*, University of Chicago Press, Chicago

Camerer C.F. (2003), *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton University Press, Princeton

Cavalli-Sforza L., Feldman M.W. (1981), *Cultural Transmission and Evolution*, Princeton University Press, Princeton, NJ

Fehr E., Fischbacher U. (2004), "Social Norms and Human Cooperation", *TRENDS in Cognitive Sciences*, 8 (4), 185-190

Gintis H. (2000), "Strong Reciprocity and Human Sociality", Journal of Theoretical Biology, 206, 169-179

Gintis H., Bowles S., Boyd R., Fehr E. (2003), "Explaining Altruistic Behavior in Humans", *Evolution and Human Behavior*, 24, 153-172

Lumsden C.J., Wilson E. O. (1981), *Genes, Mind, and Culture: The Coevolutionary Process*, Harvard University Press, Cambridge, MA

Richiardi M. (2005), *From Moral to Social Norms and Back*, LABORatorio R. Revelli Working Paper No. 38.

Simon H.A. (1993), “Altruism and economics”, *American Economic Review*, 83, 156-161.

Sober E., Wilson D.S. (1998), *Onto Others: The Evolution and Psychology of Unselfish Behavior*, Harvard University Press, Cambridge, MA

Sonnessa M. (2004), “JAS: Java Agent-based Simulation Library, an Open Framework for Algorithm-Intensive Simulations”, in Contini B., Leombruni R., Richiardi M. (eds), *Industry and Labor Dynamics: The Agent-Based Computational Economics Approach*, World Scientific, Singapore

Trivers R. L. (1971), “The evolution of reciprocal altruism”, *Quarterly Review of Biology*, 46, 35–57.

Wilson D.S., Dugatkin L.A. (1997), “Group selection and assortative interactions”, American Naturalist, 149, 336-351.